CMLS, Cell. Mol. Life Sci. 58 (2001) 72–79
1420-682X/01/010072-08 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Review

Protein folds: molecular systematics in three dimensions

C. Zhang[a] and C. DeLisi[b,*]

[a]Department of Chemistry and E. O. Lawrence Berkeley National Laboratory, University of California, Berkeley (California 94720, USA)
[b]Department of Biomedical Engineering, Boston University College of Engineering, Boston (Massachusetts 02215, USA), Fax +1 617 353 5929, e-mail: delisi@bu.edu

Received 20 June 2000; accepted 18 July 2000

Abstract. Advances in methods of structure determination have led to the accumulation of large amounts of protein structural data. Some 500 distinct protein folds have now been characterized, representing one-third of all globular folds that exist. The range of known structural types and the relatively large fraction of the protein universe that has already been sampled have greatly facilitated the discovery of some unifying principles governing protein structure and evolutionary relationships. These include a highly skewed distribution of topological arrangements of secondary-structure elements that favors a few very common connectivities and a highly skewed distribution in the capacity of folds to accommodate unrelated sequences. These and other observations suggest that the number of folds is far fewer than the number of genes, and that the fold universe is dominated by a small number of giant attractors that accommodate large numbers of unrelated sequences. Thus all basic protein folds will likely be determined in the near future, laying the foundation for a comprehensive understanding of the biochemical and cellular functions of whole organisms.

Key words. Structural genomics; motif; number of folds; classification; protein family; evolution; functions and pathways.

Introduction

More than 10,000 protein structures are now in the protein data bank (PDB) [1]. These proteins cluster into approximately 1200 sequence families, and the families in turn are distributed among some 500 folds (SCOP [2] version 1.48). Equally striking is the rate at which additional structures are being determined; e.g., the past 2 years have seen 458 new sequence families added to the PDB. Although the number of sequence families in nature is probably orders of magnitude larger than the number in the PDB, many of the families that are currently not represented will turn out to have folds that are already known. In fact, as we show below, the number of unknown folds is only twice the number of folds currently known. This relatively small number of total folds, the increasing rate at which structures are being determined, and the discovery of structural principles governing function all point toward the expectation of dramatic near-term advances in our understanding of cellular life in molecular terms.

Structural systematics

Progress in uncovering regularities in protein structure has developed in part because of the abundance of structures, in part because of the development of effective, phenetic classifications of structures [2–5]. These classifications build upon the observation that globular proteins are organized as a structural hierarchy [6]. At the base of the hierarchy are the regular secondary

* Corresponding author.

structures, e.g., α-helices and β-strands, where consecutive residues adopt similar backbone conformations. Tertiary structure is then formed by packing secondary structural elements into one or several compact globular units called domains [7]. Some proteins contain several polypeptide chains arranged in a quaternary structure.

The rigid framework formed by secondary structures is the best-defined part of a protein structure. The spatial organization of secondary structural elements, or topology, has been the primary means by which protein structures and their commonalities are characterized and classified [8, 9]. For example, the SCOP [2] database places protein domains in the same fold category if they have the same secondary structure elements in the same order, with the same topologies. Figure 1 shows the 15 most populated folds defined using this criterion. A recent comparison of SCOP with two other widely cited databases—FSSP [4] and CATH [5]—indicates that proteins assigned similar folds in one database, are generally assigned similar folds in the others [10]. The overall agreement suggests the existence of a natural logic in structural classification.

In addition to introducing order to the growing volume of structural data, the phenetic descriptions of protein structure also provide powerful clues to evolutionary relationships [11, 12]. Empirical observations support the notion that structure is more robust than sequence [13–16]. Thus, proteins that have diverged beyond significant sequence similarity still retain the three-dimensional fold of their ancestors. There are many examples where remote homology relationships that are hidden at

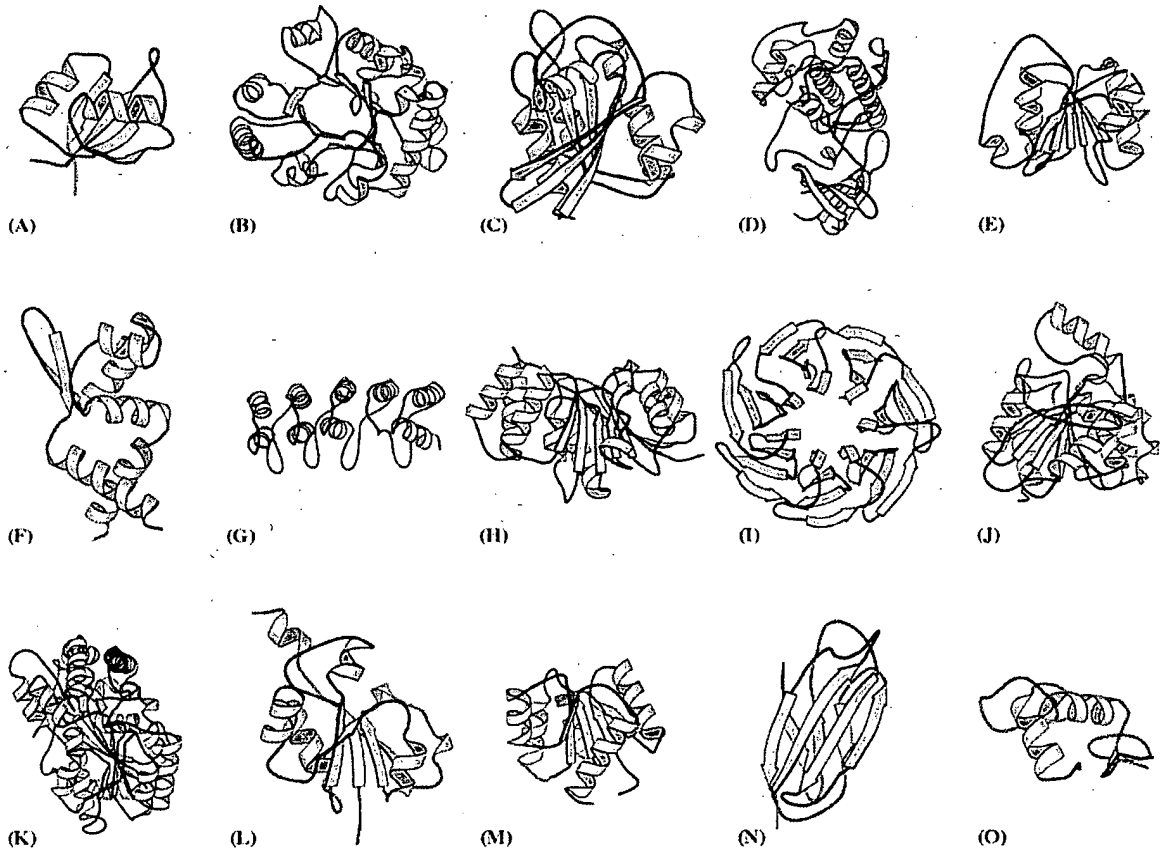

Figure 1. The 15 most populated folds. They were selected on the basis of a structural annotation of proteins from completely sequenced genomes of 20 bacteria, five Archaea, and three eukaryotes [C. Zhang, unpublished data]. From left to right and top to bottom, they are: ferredoxin-like (4.45%) (*A*), TIM-barrel (3.94%) (*B*), P-loop containing nucleotide triphosphate hydrolase (3.71%) (*C*), protein kinases (PK) catalytic domain (3.14%) (*D*), NAD(P)-binding Rossmann-fold domains (2.80%) (*E*), DNA/RNA-binding 3-helical bundle (2.60%) (*F*), α-α superhelix (1.95%) (*G*), S-adenosyl-L-methionine-dependent methyltransferase (1.92%) (*H*), 7-bladed beta-propeller (1.85%) (*I*), α/β-hydrolases (1.84%) (*J*), PLP-dependent transferase (1.61%) (*K*), adenine nucleotide α-hydrolase (1.59%) (*L*), flavodoxin-like (1.49%) (*M*), immunoglobulin-like β-sandwich (1.38%) (*N*), and glucocorticoid receptor-like (0.97%) (*O*), where the values in parentheses are the percentages of annotated proteins adopting the respective folds.

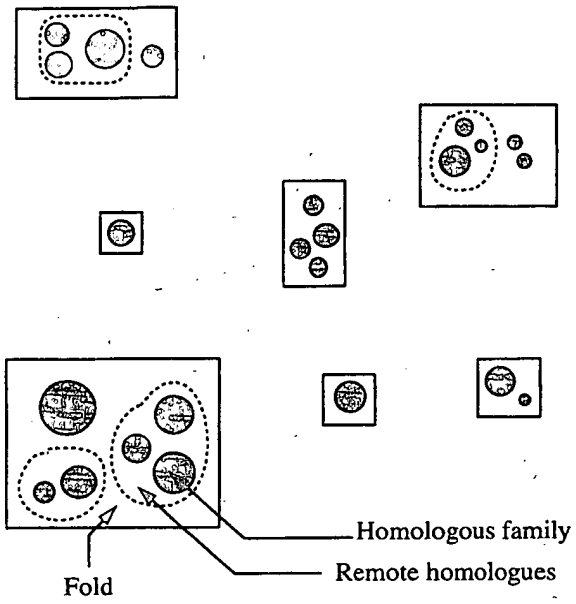

Figure 2. The nested neighborhood structure of families and folds in an abstract protein space. Protein families are represented by filled circles, folds by rectangles. Families with evidence of common ancestry are grouped into superfamilies (enclosed by dotted lines).

the sequence level have been revealed by structure comparison. When proteins are mapped onto an abstract space such that similar proteins are neighbors, proteins unified by sequence similarity (protein families) and proteins unified by structural similarity (folds) form a nested neighborhood structure (fig. 2). Some protein families that belong to the same fold may be further grouped into superfamilies based on their shared ancestry, despite low sequence similarity. Note that proteins of independent origin may well have similar structures for purely physicochemical reasons.

Known protein folds differ markedly in the number of sequence families they can accommodate [4, 17, 18]. Although a majority of the folds have only one or two representatives in the current set of structurally characterized proteins, a small number of folds are associated with many unrelated families of sequences. Below, we describe a robust estimate of the total number of folds [19] which agrees quantitatively with this observation.

How many folds are there?

To be specific, we use the SCOP (release 1.48) classification as a standard; the use of other classifications gives similar results. The breakdown of folds by the number of families in the current structural database follows a geometric distribution. More specifically, on a plot of the number of globular folds that contain m sequence families against m, the distribution drops exponentially for $m \leq 6$ (fig. 3). If we take the protein families whose structures are known as a random sample from the pool of all sequence families, the observed distribution is best explained if the distribution of protein families among folds in the universe is also geometric [19]. The total number of folds is

$$N = \frac{M_s N_s}{M_s - (1 - M_s/M)N_s} \quad (1)$$

where M_s and N_s are the observed numbers of sequence families and folds, respectively, and M is the total number of sequence families in nature.

To apply equation 1, it is important to first remove all folds with more than six sequence families because they belong to the non-exponential tail of the observed distribution. This leaves us with $N_s = 477$ folds that cover $M_s = 771$ families. Because $M \gg M_s$, the number of folds can be approximated by $N = 771 \times 477/(771 - 447) = 1250$. Figure 3 shows that the theoretical sampling distribution calculated using $N = 1250$ matches the observed distribution remarkably well. Adding back the 33 superfolds (containing 423 families) that were removed places the final estimate of N at about 1300.

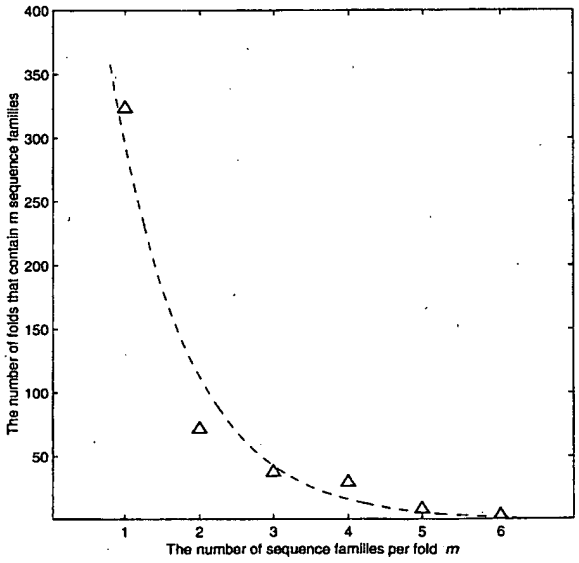

Figure 3. The breakdown of protein folds by the number of sequence families (m). Solid triangles represent the observed distributions based on the SCOP database (only the first six terms are shown). The dashed line represents the theoretical distribution based on the assumption that the sequence families whose structures have been determined are a random sample from the pool of all existing sequence families. For details, see Zhang and DeLisi [19].

Given the total number of sequence families M and the total number of folds N, the relationship between the protein families chosen for structure determination, M_s, and the fraction of folds they represent, $\Delta = N_s/N$, is given by [19]

$$\Delta = \frac{M_s}{M_s + (1 - M_s/M)N} \quad (2)$$

Because of the skewed distribution of protein families among folds, complete elucidation of all the folds by the default strategy (random family selection) still means solving the structures of virtually all non-homologous proteins, despite the fact that the number of folds in nature is limited.

As more protein structures are solved, novel folds will continue to be observed. A more practical question is how long it will take to obtain structural sketches for a majority—say, 90%—of the folds. Based on equation 2, and again assuming $M \gg M_s$, to uncover 90% of the folds requires the structural determination of representatives from 12,000 randomly chosen protein families. According to SCOP, the structures of 458 new protein sequence families have been reported over the past 2 years. If this rate continues, it will take approximately 50 years to identify 90% of the folds. To reduce this figure to 10 years, we have to increase the rate of family structural determination by about 20% per year. The alternative is to select new sequences for structural determination in accordance with a definite strategy that maximizes the chance of uncovering a new fold. Such strategies can be developed [20, 21], but their implementation will require cooperation among the community of structural biologists, at a level similar to that developed in the genomics community during the past decade. The potential payoff could be the solution to the protein-folding problem during the next decade.

Why do proteins prefer a small number of folds?

Structural regularities of proteins have long been recognized to be not only present at the whole protein (or domain) level, but also at the substructural level [6]. Secondary structure elements are observed to combine in specific geometric arrangements. The three basic supersecondary structural motifs, α-hairpin, β-hairpin, and $\beta\alpha\beta$-unit (fig. 4), are the simplest examples of such regularities [9, 22, 23]. These motifs are found more frequently in superfolds than in other folds [24], suggesting a high degree of correlation between the simplicity of secondary structure arrangement and the capacity of the fold. In general, protein structures have a tendency to place sequential structural elements adjacent in the three-dimensional space. There are indications that such placements support rapid and convenient folding.

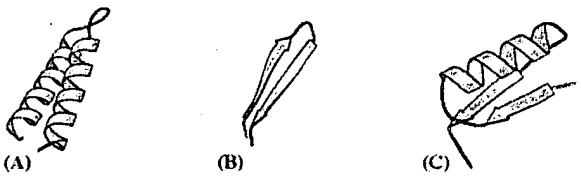

Figure 4. Supersecondary structural motifs: an α-hairpin (A), a β-hairpin (B), and a $\beta\alpha\beta$-unit (C).

For example, a significant correlation has been found between the folding rate of small proteins and the average sequence separation between contacting residues in the native state [25].

Analysis of larger substructural motifs was expected to reveal more about the topological preference of proteins, but the identification of such motifs was previously impractical because available data were limited [26]. The exception was the Greek key motif [27] which was noticed shortly after the first few β-sandwich structures had been solved [27, 28]. A Greek key motif contains four consecutive antiparallel β-strands with the first strand hydrogen bonding to the last strand (fig. 5). In particular, a composite form that consists of two

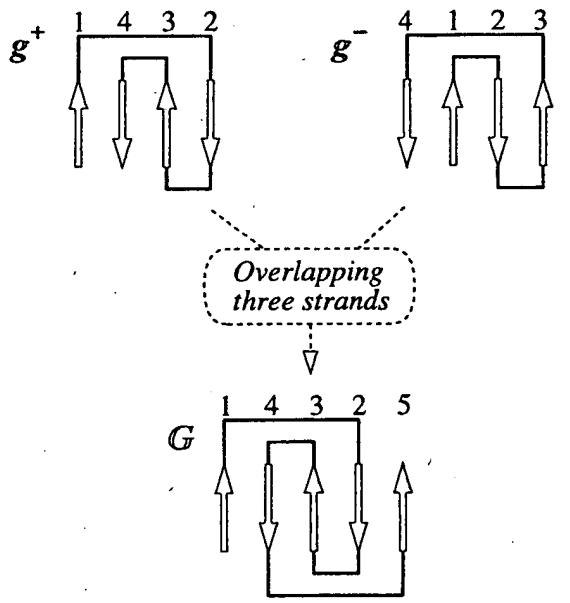

Figure 5. The Greek key motif. At the top are the two forms of the Greek key motif. A composite motif consisting of two overlapping Greek keys, shown at the bottom, is the structural determinant of β-sandwiches. For details, see Zhang and Kim [29].

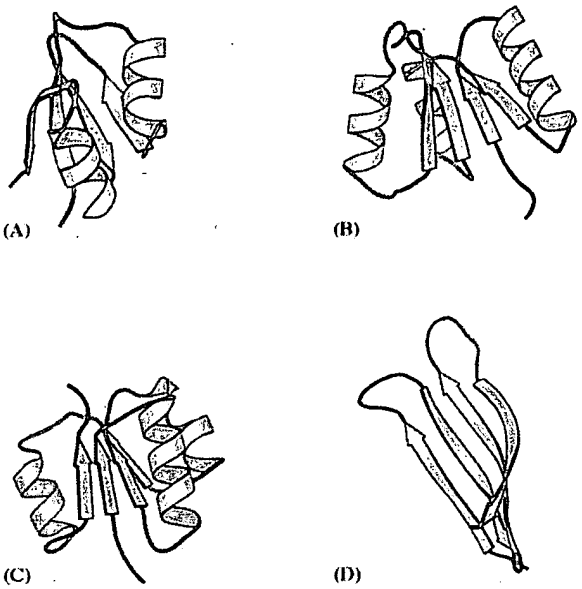

Figure 6A–D. The most common four-stranded β-sheet motifs in open-faced β-sheet structures.

overlapping Greek keys (fig. 5) is present in all known β-sandwiches [29]. Considering the enormous number of topologies that could be generated by two interwining β-sheets, the prevalence of this unit is striking. It is tempting to speculate that the unit may play an important role in the folding of β-sandwich structures.

Although the exact folding mechanisms of β-sandwiches remain unknown, recent experimental work has shown that several Ig-like β-sandwich proteins do appear to share a common folding pathway [30]. The folding nucleus residues of one of the Ig-like protein domains (FN3) were indeed mapped onto the composite Greek key motif [31]. In general, the hypothesis that the folding mechanism of a protein is dependent primarily on the topology of the native state rather than on specific details of the sequence has gained increasing support from the experimental studies of protein folding [32, 33].

With the amount of structural data currently available, our understanding of the topological preferences of substructural motifs has been greatly extended; some general principles can now be formulated. In particular, more than 50% of protein domains consist of an open-faced β-sheet flanked with helices or loops on either or both sides (see fig. 1 for examples). In these structures, the topologies of the β-sheets often determine the topologies of the entire folds. A recent survey of these structures indicates that open-faced β-sheets with more than four β-strands usually contain at least one four-stranded β-sheet substructure [34]. The β-sheet substructures thus extracted were used to analyze the topological preferences of all 96 possible four-stranded β-sheet topologies. Of the 42 topologies that have been observed, four (fig. 6) account for 50% of the open-faced β-sheet structures currently known. With the exception of the simple up-and-down meander topology (fig. 6D), the other three topologies all have at least one pair of consecutive β-strands separated by other strands. In particular, the double-stranded crossover motif (fig. 6A) has two $\beta\alpha\beta$ split crossovers [35]. As with the Greek key motif, the high frequencies of these motifs reflect an inherent bias in the natural usage of β-sheet motifs.

Most of the unobserved four-stranded β-sheet topologies fall into two groups [19]. The first group contains topologies with alternating parallel and antiparallel β-ladders (i.e., a pair of adjacent β-strands hydrogen bonded to each other). Their rare occurrence reflects the expectation that matching different hydrogen-bonding patterns is energetically unfavorable. The topologies in the second group have complex traces and may require a specific sequence of steps during folding [36, 37]; this may result in low designability [38]. Taken together, these two groups of topologies may represent a section of topological space that is not readily accessible to proteins. This indicates that we have already seen a majority of the four-stranded β-sheet topologies that exist in nature, and most of the topologies that have not been identified may have never occurred.

Larger protein structures currently known in general lack the diversity in the overall topological patterns that would be expected if the constituent secondary structural elements were arranged freely. A majority of these structures utilize recurrent substructural units as the core building blocks [9, 26, 29, 34]. Therefore, the topological biases at the substructural level directly influence the diversity of protein folds. Understanding these biases and the underlying physics helps in understanding why proteins prefer only a small fraction of the structural patterns.

Folds, functions, and pathways

The native structure is an absolute requirement for protein function. Although knowing the fold alone usually does not give definite answers to all questions regarding function [39], the rather small number of basic protein folds provides a concise and powerful framework to organize the far larger number of biological functions needed by a living cell [40]. The major route of functional evolution is local mutation. Residues change as a protein evolves to satisfy modified functional constraints, while the basic biochemical

mechanism and the overall three-dimensional fold remain unaltered. In most protein families, naturally occurring polymorphisms concentrate on residues that modulate the specificity of biological function [41]. Improvements in both efficiency and specificity through customization of the active-site architectures is the basic tenet of biological evolution. Understanding how this is achieved and compiling a comprehensive mapping between protein folds and their related functions will be a major goal of structural biologists in the next few years. Although exploring the evolution of proteins and their functions in light of structural data is only just beginning, some fundamental relationships between folds and functions have already been revealed [42–44]. For example, many analogous proteins, i.e., proteins sharing a common fold but not a common ancestor, are found to have functional sites in a common location. These locations are called supersites by Russell et al. [42]. Probably the most widely known supersite occurs within the α/β (or TIM)-barrels (fig. 1), which have long been known to bind substrates at the C-terminal end of the β-strands forming the barrel [45, 46]. Other supersites can be found in Rossmann-type doubly wound α/β folds, β-propellers, and up-and-down β-barrels [42]. Ferredoxin-like folds (fig. 1) show a tendency to bind substrates on the side of the β-sheet without α-helices packing against it. In many of the proteins adopting this fold, the β-sheet curves on the side without packing α-helices to form a concave surface where substrates bind. A common location of binding sites within analogous proteins suggests a structure-function relationship of general nature.

Elucidation of the structure and function of proteins and their interactions, and the discovery of principles that provide unity to the enormous diversity of structural data, also have a deep impact on our understanding of the complex biochemical pathways. Structure mediates biological recognition, both within and between cells. The signals impinging on the cell surface are the inputs—the boundary conditions—that modulate a complex network of interactions within the cell. The network is far more plastic than the neural net. The complement of genes that can be expressed by a cell defines the potential network, but subnets will be selected based on the specific signals on the surface as well as the biochemical environment inside the cell. Much of the network connectivity, the links, is mediated by interactions between proteins. Complex systems analysis [47] teaches us that, depending on its topology (connectivity), the qualitative behavior of a network can change (e.g., a different group of genes could be induced, the cell fate could be altered). In this way, the behavior of the cell can be seen to be qualitatively dependent on the local structure of one or more proteins. Selectively targeting these molecules based on knowledge of their structures would provide a way to control cell behavior, an approach that will reach its full power when a sufficient number of structures are available.

Concluding remarks

The first protein fold classification, guided by the visual recognition of recurrent folding patterns, dates back to the late 1970s [8, 9]. This work has been taken much further recently, with several systems available that provide comprehensive classifications of all experimentally determined structures. In parallel, many automatic procedures have been developed that recognize structural similarities between proteins [3]; some of these procedures [48–50] are now used routinely to compare a newly solved protein structure with structures in the PDB. This bevy of structural bioinformatic tools has been used to infer ancient evolutionary relationships [11, 12] and to suggest functional mechanisms for hypothetical proteins [51].

Fold classifications are only the first step toward a global and comprehensive understanding of protein folds. To reveal the principles underlying the design of protein folds and find answers to many other fundamental questions in structural biology requires a deeper understanding of the relationships among protein folds. The focus will be the high-order substructural motifs that are the common building blocks of many protein folds. These motifs organize the fold space into distinct attractors. The available data show that perhaps a majority of these motifs have already been observed in the known protein structures [26, 29, 34, 52]. More sensitive methods for recognizing such motifs will be of great value. Given the importance of topology in determining the protein-folding mechanism, complete knowledge of the core folding units will facilitate the development of more effective methods for protein structure prediction.

The estimates made here and elsewhere [19, 53–56] suggest that there may be a limited number of folds available to proteins. However, because of the skewed distribution of proteins among folds, the effort to completely elucidate all existing folds will benefit greatly from a definite strategy that can maximize the information return from experimental structure determination. A joint sequence and structural classification of protein families offers powerful clues for judiciously choosing novel targets [20, 21, 57–60].

Structural information is becoming an indispensable component of our understanding of a variety of biological phenomena. As more structures are determined, our understanding of how function is modulated by sequence changes will improve. Molecular systematics based on protein structure provides an effective way to

organize the large body of functional data and to search for unified principles. In a foreseeable future, emergent cell properties and their control by human intervention will be traced directly to protein structure and its modulation.

1 Bernstein F. C., Koetzle T. F., Williams G. J. B., Meyer E. F. Jr, Brice M. D., Rodgers J. R. et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. **112:** 535–542
2 Murzin A., Brenner S. E., Hubband T. and Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247:** 536–540
3 Wodak S. J. (1996) Extending molecular systematics to the third dimension. Nat. Struct. Biol. **3:** 575–578
4 Holm L. and Sander C. (1996) Mapping the protein universe. Science **273:** 595–602
5 Orengo C. A., Michie A. D., Jones S., Jones D. T., Swindells M. B. and Thornton J. M. (1997) CATH—a hierarchic classification of protein domain structures. Structure **5:** 1093–1108
6 Branden C. and Tooze J. (1999) Introduction to Protein Structure. Garland, New York
7 Wetlaufer D. B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. USA **70:** 697–701
8 Levitt M. and Chothia C. (1976) Structural patterns in globular proteins. Nature **261:** 552–558
9 Richardson J. S. (1981) The anatomy and taxonomy of protein structure. Adv. Protein Chem. **34:** 167–339
10 Hadley C. and Jones D. T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure Fold Des. **7:** 1099–1112
11 Murzin A. G. (1998) How far divergent evolution goes in proteins. Curr. Opin. Struct. Biol. **8:** 380–387
12 Holm L. (1998) Unification of protein families. Curr. Opin. Struct. Biol. **8:** 372–379
13 Chothia C. and Lesk A. M. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J. **5:** 823–826
14 Sander C. and Schneider R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct. Funct. Genet. **9:** 56–68
15 Rost B. (1997) Protein structures sustain evolutionary drift. Fold. Des. **2:** S19–S24
16 Wood T. C. (1999) Evolution of protein sequences and structures. J. Mol. Biol. **291:** 977–995
17 Orengo C. A., Jones D. T. and Thornton J. M. (1994) Protein superfamilies and domain superfolds. Nature **372:** 631–634
18 Brenner S. E., Chothia C. and Hubbard T. J. P. (1997) Population statistics of protein structures: lessons from structural classifications. Curr. Opin. Struct. Biol. **7:** 369–376
19 Zhang C. and DeLisi C. (1998) Estimating the number of protein folds. J. Mol. Biol. **284:** 1301–1305
20 Gaastraland T. (1998) Structural genomics: bioinformatics in the driver's seat. Nat. Biotechnol. **16:** 625–627
21 Terwilliger T. C., Waldo G., Peat T. S., Newman J. M., Chu K. and Berendzen J. (1998) Class-directed structure determination: foundation for a protein structure initiative. Protein Sci. **7:** 1851–1856
22 Taylor W. R. and Thornton J. M. (1984) Recognition of super-secondary structure in proteins. J. Mol. Biol. **173:** 487–512
23 Sibanda B. L. and Thornton J. M. (1985) Hairpin families in globular proteins. Nature **316:** 170–174
24 Salem G. M., Hutchinson E. G., Orengo C. A. and Thornton J. M. (1999) Correlation of observed fold frequency with the occurrence of local structural motifs. J. Mol. Biol. **287:** 969–981
25 Plaxco K. W., Simons K. T. and Baker D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. **277:** 985–994
26 Efimov A. V. (1994) Common structural motifs in small proteins and domains. FEBS Lett. **355:** 213–219
27 Richardson J. S. (1977) β-Sheet topology and the relatedness of proteins. Nature **268:** 495–500
28 Ptitsyn O. B. and Finkelstein A. V. (1980) Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? Q. Rev. Biophys. **13:** 339–386
29 Zhang C. and Kim S-H. (2000) A comprehensive analysis of the Greek key motifs in β-barrels and β-sandwiches. Proteins Struct. Funct. Genet. **40:** 409–419
30 Clarke J., Cota E., Fowler S. B. and Hamill S. J. (1999) Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. Structure **7:** 1145–1153
31 Hamill S. J., Steward A. and Clarke J. (2000) The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. J. Mol. Biol. **297:** 165–178
32 Alm E. and Baker D. (1999) Matching theory and experiment in protein folding. Curr. Opin. Struct. Biol. **9:** 189–196
33 Goldenberg D. P. (1999) Finding the right fold. Nat. Struct. Biol. **6:** 987–990
34 Zhang C. and Kim S-H. (2000) The anatomy of protein β-sheet topology. J. Mol. Biol. **299:** 1075–1089
35 Orengo C. A. and Thornton M. J. (1993) Alpha plus beta folds revisited: some favoured motifs. Structure **1:** 105–120
36 Onuchic J. N., Luthey-Schulten Z. and Wolynes P. G. (1997) Theory of protein folding: the energy landscape perspective. Annu. Rev. Phys. Chem. **48:** 545–600
37 Dill K. A. and Chan H. S. (1997) From Levinthal to pathways to funnels. Nat. Struct. Biol. **4:** 10–19
38 Li H., Helling R., Tang C. and Wingreen N. (1996) Emergence of preferred structures in a simple model of protein folding. Science **273:** 666–669
39 Martin A. C., Orengo C. A., Hutchinson E. G., Jones S., Karmirantzou M., Laskowski R. A. et al. (1998) Protein folds and functions. Structure **6:** 875–884
40 Thornton J. M., Orengo C. A., Todd A. E. and Pearl F. M. G. (1999) Protein folds, functions and evolution. J. Mol. Biol. **293:** 333–342
41 Parham P., Lomen C. E., Lawlor D. A., Ways J. P., Holmes N., Coppin H. L. et al. (1988) Nature of polymorphism in HLA-A, -B, and -C molecules. Proc. Natl. Acad. Sci. USA **85:** 4005–4009
42 Russell R., Sasieni P. and Sternberg M. (1998) Supersites within superfolds: binding similarity in the absence of homology. J. Mol. Biol. **282:** 903–918
43 Koonin E. V., Tatusov R. L. and Galperin M. Y. (1998) Beyond complete genomes: from sequence to structure and function. Curr. Opin. Struct. Biol. **8:** 355–363
44 Orengo C. A., Todd A. E. and Thornton J. M. (1999) From protein structure to function. Curr. Opin. Struct. Biol. **9:** 374–382
45 Farber G. K. and Petsko G. A. (1990) The evolution of α/β barrel enzymes. Trends Biochem. Sci. **15:** 228–234
46 Hegyi H. and Gerstein M. (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J. Mol. Biol. **288:** 147–164
47 Kauffman S. A. (1993) The Origins of Order. Oxford University Press, New York
48 Holm L. and Sander C. (1993) Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233:** 123–138
49 Gibrat J. F., Madej T. and Bryant S. H. (1996) Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. **6:** 377–385

50 Shindyalov I. N. and Bourne P. E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. **11:** 739–747
51 Zarembinski T. I., Hung L-W., Mueller-Dieckmann H-J., Kim K-K., Yokota H., Kim R. et al. (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. Proc. Natl. Acad. Sci. USA **95:** 15189–15193
52 Chothia C., Hubbard T., Brenner S., Barns H. and Murzin A. (1997) Protein folds in the all-β and all-α classes. Annu. Rev. Biophys. Biomol. Struct. **26:** 597–627
53 Chothia C. (1992) One thousand families for the molecular biologist. Nature **357:** 543–544
54 Wang Z-X. (1998) A re-estimation for the total numbers of protein folds and superfamilies. Protein Eng. **11:** 621–626
55 Govindarajan S., Recabarren R. and Goldstein R. A. (1999) Estimating the total number of protein folds. Proteins Struct. Funct. Genet. **35:** 408–414
56 Wolf Y. I., Grishin N. V. and Koonin E. V. (2000) Estimating the number of protein folds and families from complete genome data. J. Mol. Biol. **299:** 897–905
57 Frishman D. and Mewes H. W. (1997) PEDANTic genome analysis. Trends Genet. **13:** 415–416
58 Sali A. (1998) 100,000 protein structures for the biologist. Nat. Struct. Biol. **5:** 1029–1032
59 Wolf Y. I., Brenner S. E., Bash P. A. and Koonin E. V. (1999) Distribution of protein folds in the three superkingdoms of life. Genome Res. **9:** 17–26
60 Teichmann S. A., Chothia C. and Gerstein M. (1999) Advances in structural genomics. Curr. Opin. Struct. Biol. **9:** 390–399